LINKgroup

Distance Learning System

scikit learn

Biblioteka sklearn

Python Data Access

Šta je scikit-learn

https://scikit-learn.org/stable/

- Sklearn je biblioteka za predviđanje podataka
- Bazirana je na numpy biblioteci
- Otvorenog je koda i besplatna za korišćenje

pip install scikit-learn

Uzorci, svojstva i klase

- Sklearn radi sa podacima podeljenim na uzorke i svojstva
- Uzorci su različiti primerci neke klase (u slučaju klasifikacije)
- Svojstva su karakteristike jednog primerka neke klase
- Mogu biti **diskretna** i **numerička**

Uzorak 1

Kapa: plava
Kosulja: plava

Uzorak 2

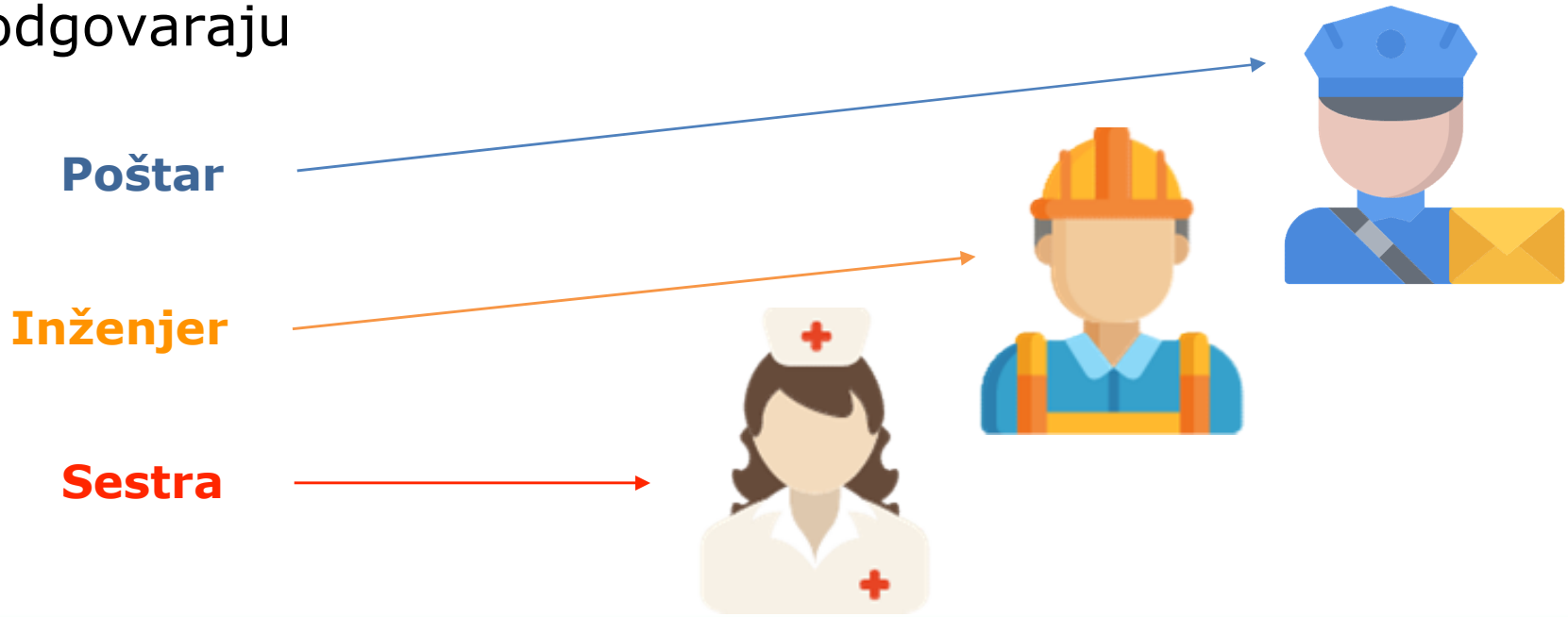
Kapa: narandzasta
Kosulja: tirkizna

Uzorak 3

Kapa: bela
Kosulja: bela

Klase

- Ukoliko se radi o klasifikaciji, osim samih podataka (uzoraka) postoje i klase kojima određeni uzorci odgovaraju



Postavljanje svojstava

- Svojstva se predstavljaju numeričkim vrednostima, a jedan uzorak nizom svojstava

Oznake

kapa = 0
kosulja = 1

Boje

plava = 1
bela = 2
narandzasta = 3
tirkizna = 4

uzorak = [1,1]

Vektorizacija svojstava

- Vektorizacija je pretvaranje kategoričkih vrednosti u numeričke nizove

Dokumenti

0 dobar dan
1 kako ste
2 dan je dobar

Reči

0	**1**	**2**	**3**	**4**
dan	dobar	je	kako	ste

Niz:

[[1,1,0,0,0] , [0,0,0,1,1] , [1,1,1,0,0]]

Vektorizacija rečnika

```
data =
[{"hero":"fiddle","damage":20},{"hero":"akali","dama
ge":30},{"hero":"fiddle","damage":50}]
```

Svojstva

0 damage

1 hero is akali

2 hero is fiddle

Niz:

[[20 , 0 , 1] , [30 , 1 , 0] , [50 , 0 , 1]]

Kreiranje i popunjavanje modela

U zavisnosti od grupe algoritama učitava se odgovarajući sklearn modul

U sklearn, modeli se nazivaju **estimatori**

```
from sklearn import linear_model

model = sklearn.linear_model.LinearRegression()
```

Metod **fit** prihvata niz uzoraka (x) i niz klasa (y) i na osnovu njih generiše model. Ovaj metod "trenira" model sa odredjenim podacima

```
x = np.array([[1],[2],[3]])
y = np.array([2,4,6])
model.fit(x,y)
print(model.coef_)
```

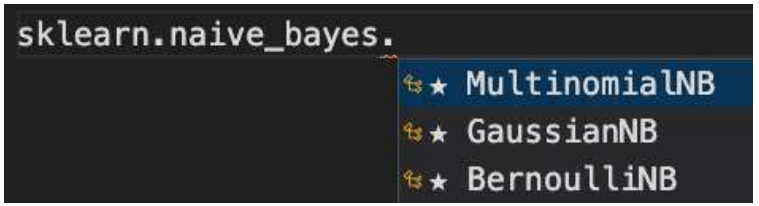
```
[2.]
```

Svojstvo **coef_** sadrži izračunati koeficijent ukoliko postoji u modelu

Dostupni modeli

Modeli (estimatori) podeljeni tematski po sklearn modulima:

- **sklearn.linear_model**
- **sklearn.svcp**
- **sklearn.naive_bayes**
- **sklearn.tree**
- **sklearn.ensamble**
- **sklearn.neural_network**



Predikcija

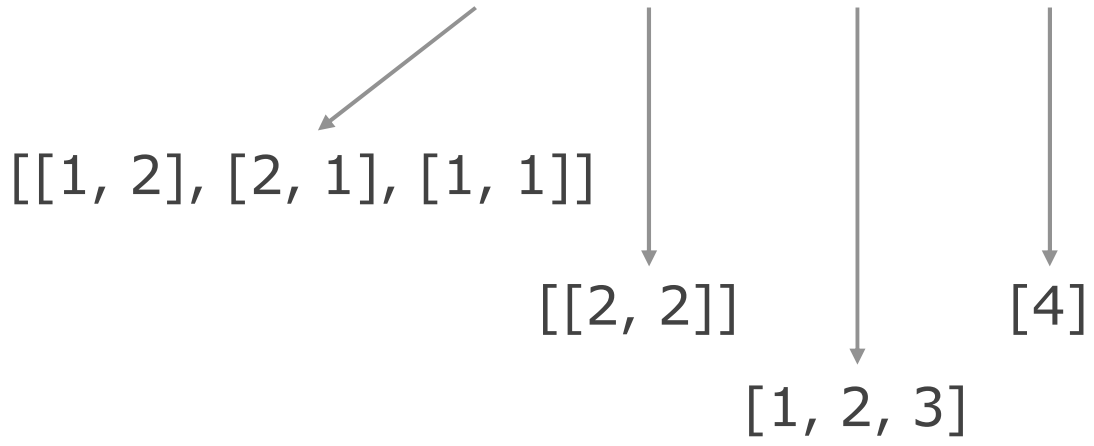
- Svaki model ima metod predict, koji prihvata kolekciju ulaznih uzoraka (nezavisne podatke) i vraća listu predviđanja

```
model = LogisticRegression()
model.fit([[1,2],[2,2],[2,1],[1,1]],[1,2,3,4])

res = model.predict([[2,2]])
```

[2]

Podela podataka modela za unakrsnu validaciju

```
from sklearn.model_selection import train_test_split

data = [[1,2],[2,1],[1,1],[2,2]]
target = [1,2,3,4]
x_train, x_test, y_train, y_test = train_test_split(data,target)
```



Provera kvaliteta algoritma

- Dobijeni rezultat se može proveriti ručno ili ugrađenim metodama
- Mnoštvo metoda za proveru kvaliteta rezultata nalazi se u paketu **metrics**

```
import sklearn.metrics as metrics
```

```
print(
    metrics.classification_report(y_test,res)
)
print(
    metrics.confusion_matrix(y_test,res)
)
```

	precision	recall	f1-score	support
0	1.00	1.00	1.00	16
1	1.00	0.90	0.95	10
2	0.92	1.00	0.96	12
accuracy			0.97	38
macro avg	0.97	0.97	0.97	38
weighted avg	0.98	0.97	0.97	38

```
[[16  0  0]
 [ 0  9  1]
 [ 0  0 12]]
```

Provera kvaliteta regresije

```
model = LinearRegression()
model.fit(x_tr,y_tr)
pred = model.predict(x_tst)
print(
    metrics.explained_variance_score(y_tst,pred)
)
```

Rad sa podacima za vežbanje

https://scikit-learn.org/stable/datasets/index.html

- Scikit learn sadrži pakete podataka za vežbanje ili produkcionu upotrebu
- Deo paketa je automatski pridružen biblioteci, dok je deo potrebno eksplicitno učitati (downloadovati)
- Pridruženi paketi se učitavaju **load** funkcijama modula **datasets**
- Preuzimanje većih setova podataka, obavlja se **fetch** funkcijama

load_boston
load_iris
load_diabetes
load_linnerud
load_wine
load_brest_cancer

fetch_olivetti_faces
fetch_20newsgroups
fetch_california_housing
fetch_covtype
fetch_kddcup99

fetch_lfw_pairs
fetch_lfw_people
fetch_openml
fetch_rcv1
fetch_species_distributions

Credits

https://www.flaticon.com/authors/freepik https://www.flaticon.com/authors/nikita-golubev